Genfam: integrative system for gene family analysis,
including a method of evolutionary event identification and evidences for an involvement in environmental stress response

Delphine Larivière*, Jonathan Lorenzo*, Stéphanie Sidibe-Bocs*, Jean-François Dufayard*, and Dominique This**
* CIRAD, UMR AGAP ; ** Montpellier SupAgro, UMR AGAP:
TA-A 108/03. Avenue Agropolis. 34398 Montpellier cedex, France
delphine.lariviere@cirad.fr, jonathan.lorenzo@cirad.fr, stephanie.sidibe-bocs@cirad.fr, jean-francois.dufayard@cirad.fr, dominique.this@supagro.fr

UMR AGAP - CIRAD "actions incitatives"

Drought Conference - June 2015

Abstract - *Important research efforts are made to characterize mechanisms of biological interest, such as stress tolerance, through gene family studies. The identification of these families allows the functional annotation of genes, as genes belonging to one family are supposed to have similar or related functions. We have developed an online comparative genomics application, GenFam, allowing the user to build custom gene families based on several data sources, to run analysis workflows, and to gather results into a synthetic visualization. It allows displaying functional evidences and evolutive history information through widely used tools and the new algorithm IDEVEN. This algorithm is complementing phylogenetic analyses by using syntenic block data and synonymous mutation rates (dS). The system contains additional modules oriented towards stress response study, such as (differential) gene expression data, functional annotations (gene ontologies), and the identification of specific stress-related cis elements in the promoter regions. The aim of this work is to facilitate knowledge representation and functional inference for scientists working on gene families, and an easier way to link stress-related evidences within a gene family tree. It can therefore highlight adaptive evolution within gene families in relation to stress-prone environments and identify candidate genes for drought tolerance in non-model crops by translational studies.*

Gene family composition and analysis - the first GenFam workflow

Small sequence file (<20 sequences)

Sequences of interest - Homologous genes
Small number of sequences, carefully chosen

Searching for homologous genes in a sample of species

HMMer : HMM Build / HMM Search
(e.g. Coffea + Grapevine + Rice …)

Sequence file (whole family)

Galaxy

Aligning sequences

MAFFT + (GBlocks or TrimAl)

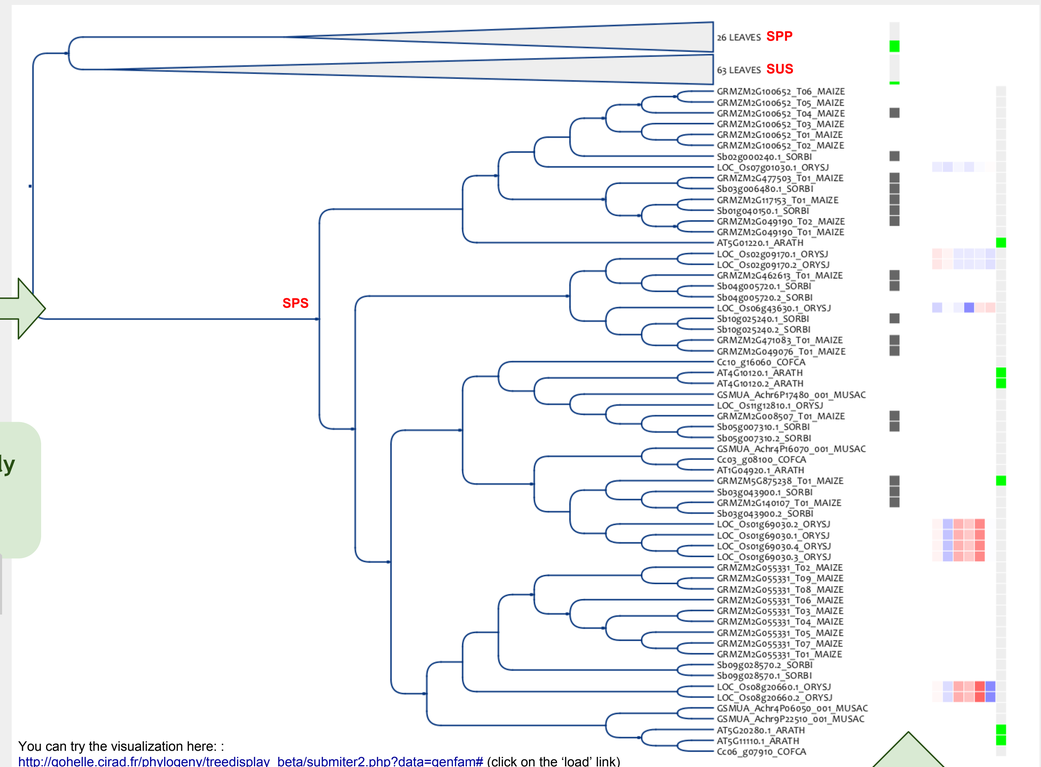
Multiple alignment

Phylogenetic analysis

PhyML (tree building)
RapGreen (duplication inference)

Tree + Orthology / paralogy

Phylogenetic tree representation

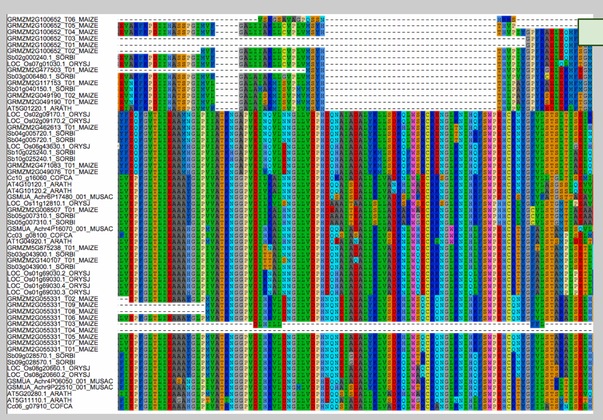


You can try the visualization here: :
http://gohelle.cirad.fr/phylogeny/treedisplay_beta/submiter2.php?data=genfam# (click on the 'load' link)

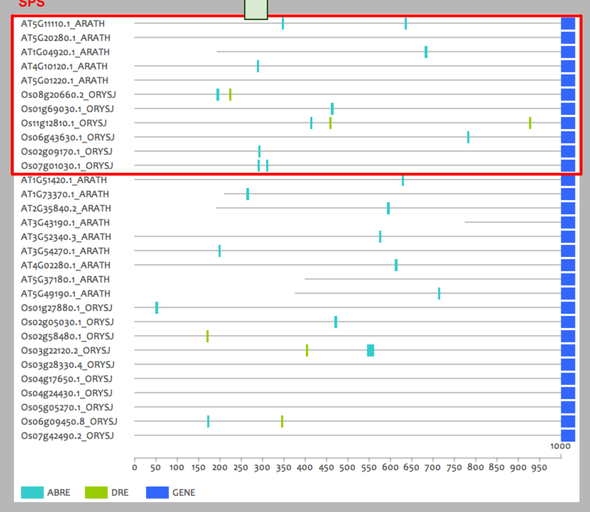
Data integration around the gene family

Visualization dashboard

Example of SPS family



Alignment



Promoter analysis

Get promoters

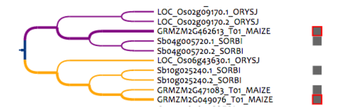
Specific developments

Detection of binding sites (TFBS)

(e.g. PlantPAN)

Heterogeneous data

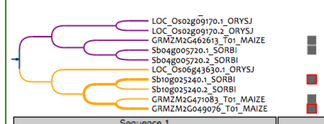
Sequence 1	Sequence 2	Relationship	Score
GRMZM2G462613_T01_MAIZE	GRMZM2G049076_T01_MAIZE	WGD	0.8186



Ideven

Detection of syntenic evidences about orthology and paralogy

CoGe
Accelerating Comparative Genomics



Sequence 1	Sequence 2	Relationship	Score
Sb10g025240.1_SORBI	GRMZM2G049076_T01_MAIZE	ortholog	0.1036

GEO Gene Expression Omnibus

Coffee Genome Hub
a genomics and genetics resource for coffee

Expression data gathering

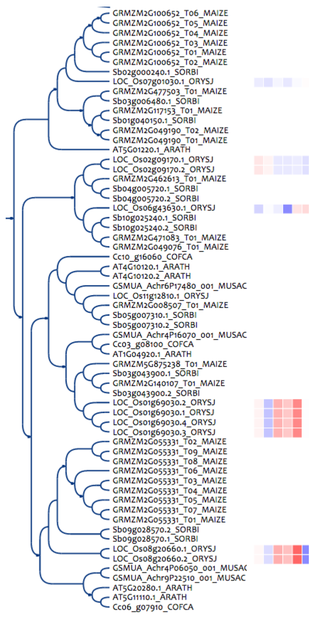
Get tissular and differential expression data

Expression data display

Display expression data in the phylogenetic visualization

Expression data analysis

Discuss the evolution of expression profile along the evolutionary history



Annotation data

UniProt

GMOD

The Banana Genome Hub

. . .

Annotation gathering

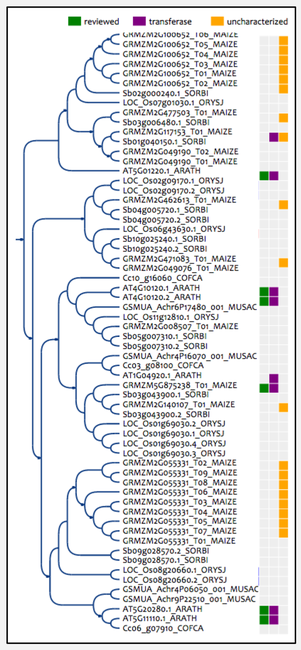
Collect from heterogeneous databases
(Chado/GMod, UniProt…)

Standardization

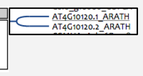
Unit the annotation in a single table

Display

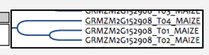
Display in the tree, with popup,
and colored squares on front of leaves



Sequence	Code	Function	Reviewed	Score
AT4G10120.1_ARATH	SPS4	Probable sucrose-phosphate synthase 4 EC 2.4.1.14 Sucrose phosphate synthase 4F AtSPS4F UDP-glucose-fructose-phosphate glucosyltransferase	reviewed	4/5



Sequence	Code	Function	Reviewed	Score
GRMZM2G152908_T04_MAIZE	NA	Sucrose synthase1 Uncharacterized protein	no	1/5



Conclusion

In this example, subfunctionalization following old duplication events is evidenced in the SPS gene family, in relation to drought induction. However data are still incomplete.

This on-going project will provide a toolbox facilitating translational biology, from model plants (Arabidopsis, rice) to newly sequenced crops (coffee, banana).
The evolutionary history of gene families (duplications and speciation events) - carefully estimated from phylogenetic and synteny analysis - together with functional evidences (curated annotations, expression profiles, promoter structure…), all of this presented on a synthetic dashboard, provide candidates in non-model species for genes of interest in stress tolerance, as far as functional orthology can be confirmed.